动态主题网络视角下的突破性创新主题识别：以区块链领域为例*

■ 陈虹枢　宋亚慧　金茜茜　汪雪锋
北京理工大学管理与经济学院　北京 100081

摘　要：［目的/意义］突破性创新对科技发展具有关键作用。大数据环境下，科学技术发展本身所具有的复杂、多维、不断进化等特征越发凸显。以动态视角进行突破性创新主题识别，对于为国家、企业及高校详析突破性创新领域、合理配置创新资源以及提供创新升级解决方案具有重要意义。［方法/过程］综合运用主题模型、词嵌入算法以及复杂网络分析等方法构建动态主题网络，全面考量主题在时间窗口内的结构特性以及时间窗口间的演化状态，并以其为基础结合突破性创新的新颖性、突变性、影响力和学科交叉性特征识别突破性创新主题。［结果/结论］面向区块链领域展开实证研究，识别出神经网络（Neural Network）和边缘计算（Edge Computing）两个主题的突破性创新特征最为显著。结合区块链现有研究及美国国家科学技术委员会发布的关键和新兴技术清单，验证了本文方法的可行性和有效性。但有关结果的定量验证，以及融合多源数据的突破性创新主题识别有待进一步研究。

关键词：突破性创新　主题网络　主题识别　LDA　Word2vec 模型　区块链
分类号：G250.2
DOI：10.13266/j.issn.0252－3116.2022.10.004

全球新一轮科技革命蓄势待发，我国进入“十四五”发展的重要时期，习近平总书记曾多次强调，“创新是推动一个国家、一个民族向前发展的重要力量”。突破性创新（Radical Innovation）作为一种极具革命性的创新活动，是企业革新产业链条、提高竞争力的关键要素，是新时期在日趋激烈的国际竞争中把握先机的重要保障[1－3]。在“提升创新体系效能”大背景下，及时准确地识别突破性创新，是面向国家政策制定、企业战略布局、学界科研规划提供决策支持的关键一环，已经成为学术界与产业界共同关注的重要研究问题之一。

突破性创新的概念基于熊彼特提出的“创造性的破坏”[4]。W. J. Abernathy 等将其定义为利用技术创新提升企业地位、重构市场格局的创新，为后续突破性创新的研究奠定了基础[5]。作为一种非渐进式的创新活动，突破性创新具有突变性、新颖性、学科交叉性等多种特征，目前已有大量研究采用文献计量、文本挖掘以及网络分析等方法，对突破性创新识别展开研究，并取得了一定成果[1，6－8]。具体地，基于引文分析和共词分析的研究，分别从文献或专利被引数量、引文数量、引文新颖性、引文关键词或共词网络以及词频变化等角度构建相关指标识别突破性创新[9－10]。但引文分析存在一定的时滞性问题，共词分析则在探究文本语义与特征表达上存在不足。针对上述问题，已有学者结合文本挖掘和网络分析的方法进行突破性创新识别[11－12]。但是，在网络视角下充分考量技术演化的动态性以及全面测度突破性创新的多种特征等具体问题上，仍然缺乏较为系统的工作。基于已有研究，本文聚焦以下两个研究问题：①如何合理抽取并向量化主题，构建动态主题网络，反映目标领域主题演化过程与态势？②如何在问题①的基础上测度突破性创新的多种特征，更加系统地识别领域内的突破性创新主题？

主题抽取（Topic Extraction）是突破性创新主题识别的关键基础之一[13]。技术主题的抽取质量会影响

* 本文系国家自然科学基金青年项目“多源异质网络视角下产学研合作产生机理及潜在机会发现研究”（项目编号：72004009）和北京理工大学优秀青年教师学术启动项目“基于主题模型及深度学习的技术演化路径识别研究”研究成果之一。
作者简介：陈虹枢，助理教授，博士，E-mail：Hongshu.Chen@bit.edu.cn；宋亚慧，硕士研究生；金茜茜，博士研究生；汪雪锋，教授，博士。
收稿日期：2021－11－21　修回日期：2022－02－16　本文起止页码：45－58　本文责任编辑：王传清

后续主题多种性质的识别。从宏观、浅层的统计研究过渡到具体、深入的知识发现是大数据背景下科学计量学与科技文本挖掘相关方法揭示学科知识结构的研究趋势。在已有的主题抽取研究中,以关键词为基础的方法对于技术概念的表达最为细致,但往往需要进行多层、大量、有监督的筛选,筛选原则及聚类的粒度直接影响技术主题的生成,给后续主题语义表达带来了更多的挑战[14]。以隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)为代表的主题模型,由于能够深入挖掘大量文本中的隐含语义,近年来受到主题识别[15-16]、技术预测[17]、科学图谱[18]等领域学者的广泛关注,但已有研究尚未对如何合理预设主题数目达成普遍共识。同时,已有综述研究指出现有突破性创新识别方法在考量主题演化方面仍然存在局限性[13]。为了揭示科技创新过程中主题的产生、发展、演变、消亡等过程,进而更好地识别突破性创新主题,需要合理计算主题在单一时间窗口或多个时间窗口的相似度。而以 LDA 为代表的主题模型很难测度主题之间的“距离”,在计算主题相似度方面存在固有问题[19],主题识别与后续的技术演化、特征分析缺乏系统衔接。词嵌入(Word Embedding)算法可以在考虑内容上下文的同时发现大规模文本数据中的潜在语义[20]。近年来,因其将词映射到向量空间的出色能力引起了广泛的关注。词向量可用于替换科学文本挖掘中的传统单词表示,从而为主题提取及主题相似度计算带来了全新视角[21-22]。

基于以上背景,面向突破性创新识别在大数据环境下进行主题提取、关系表示及指标体系构建的全新挑战,本文综合运用主题模型、词嵌入以及复杂网络分析等方法,构建动态主题网络同时揭示主题演化过程,并以其为基础结合突破性创新多项特征对突破性创新主题进行识别。具体地,本文以多个时间窗口下的科研论文数据为数据源,综合运用概率主题模型与词嵌入的方法进行主题的抽取与向量化,克服以关键词为核心的主题识别方式在语义表达上存在盲点,以及筛选及降维困难等问题,同时完成科技文本到数学向量的映射。在连续时间窗口下,本文构建起统一向量空间内动态变化的多个主题网络,并对目标文本集在多个网络上的主题演化情况进行定量的表达与总结。最后,在分析动态主题网络的结构特性变化及知识流动的基础上,本文通过对突破性创新内涵与特征的梳理,构建起测度主题“新颖性”“突变性”“影响力”和“学科交叉性”的层次指标体系,对突破性创新主题进行识别。

1 研究现状

1.1 突破性创新内涵及特征

迄今为止,已有大量学者从不同角度、不同方面对突破性创新进行定义并展开研究,其研究维度主要包含微观与宏观两个层面。微观层面从技术本身出发,关注技术自身所带来的突破,认为突破性创新不同于渐进式创新对现有技术的微小改变和调整,而是整合新的学科知识,基于不同的科技原理,结合科学前沿与新兴技术突破现有的技术枷锁,创造革命性的技术变革[2, 23-24]。宏观层面则从创新活动所产生的实质性影响力进行定义[25],主要包括两个方面:其一是对市场或行业格局产生的影响力[24, 26];其二是在科学研究中产生的学术影响力[27]。

已有研究归纳突破性创新内涵的切入点与侧重点各有不同。本文通过系统梳理,较为全面地总结目前研究中的突破性创新特征,包含前沿性、突变性、高影响力、学科交叉性、不连续性和非线性、长期性、不确定性和不可预测性、发散性以及随机性和偶然性,具体的特征解释如表 1 所示。虽然已有研究归纳了突破性创新的多种特征,但在实际识别过程中,因其中部分特征,如不可确定性、发散性、偶然性等,难以被直接量化,在现有的定量研究中,新颖性[10, 28]、学科交叉性[29-31]、突变性[32-33]以及影响力[34]等是进行突破性创新主题识别的主要特征。因此,汲取现有研究对突破性创新内涵的理解以及主要特征选取经验,本文以主流研究中最为常用的特征为基础,即新颖性、突变性(即重大突破)、高影响力以及学科交叉性,构建层次指标体系,展开定量研究。

1.2 主题抽取与演化分析

主题抽取,即主题识别,作为文本挖掘的一项具体应用,在目前的国内外研究中,主要基于关键词(主题词)聚类、SAO 语义结构识别[45]以及概率主题模型等技术方法。总体来说,三种方法在核心技术内容的提取与表达上各有利弊:①传统的以关键词为基础的方法对于技术概念的表达最为细致,但对语义的表达则有限,且需要进行多层、大量、有监督的筛选,筛选原则及聚类的粒度直接影响技术主题的生成[14];②相较关键词,SAO 语义结构能识别语境、提升语义的表达,但在大数据环境中,以 SAO 结构为核心的方法存在降维的困难[46];③以 LDA 为代表算法的主题模型能够挖掘大量文本中的隐含语义,且以词分布(可视为词簇)的形式来表达概念可以避免同义词带来的歧义, 因而在

表 1　突破性创新特征归纳

特征	特征解释
新颖性/前瞻性/前沿性	与新兴技术主题的新颖性异曲同工，代表着前沿的技术发展与进步[8, 10, 35-36]
突变性	突破性创新标志性特征之一。突破性创新的发生是非渐进式的，技术变革强度大，往往出现了大幅度的创新或程度较大的进步性改变[8, 36-38]
高影响力	突破性创新是其他技术、产品以及服务等的基础，通常会对市场或行业格局以及科学研究产生重要影响力[10, 25, 27]
学科交叉性	突破性创新通常融合不同的科学技术原理，建立在全新的知识基础之上，是多学科交互作用、众多知识领域重组的结果，特别是通常不发生相互联系的知识领域之间发生了重组，更可能产生突破性创新[8, 33, 39-40]
不连续性和非线性	相对于渐进性创新、连续性创新而言，突破性创新在技术或市场、产品、商业模式等发生的变化是不连续的、非线性的，会发生出现－消亡－再出现的反复过程，技术轨道在演化过程中发生不连续跳跃[41-43]
长期性	突破性创新一般周期长，需要长期的培育过程，所需平均完成时间一般为 10 年以上[36, 42, 44]
不确定性和不可预测性	突破性创新的发生在技术、市场、资源及组织等多方面具有不确定性，并且其往往在新的技术轨道上发展，难以事前识别，也很难预测技术的发展方向[42, 44]
发散性	主要表现为思想产生的发散性，不是遵循原有的方法与路径[42]
随机性和偶然性	突破性创新的产生并不是预设好的，其往往产生于许多偶然和随机出现的新思想[42]

近 10 年中被广泛用于主题抽取研究。但是，由于 LDA 的主题总数作为参数需要提前设置[47]，且过大或过小都会影响主题抓取及表达的准确度和可读性，虽然已有研究形成通过困惑度（Perplexity）确定主题数量的解决方案[48]，但在实际研究中往往需要学者们继续对主题进行评价或筛选来平衡主题的可读性[48]。

在主题抽取之后，揭示科研主题的演化过程、规律和态势对于把握领域发展趋势以及突破性创新主题的探测均具有重要意义。早在 20 多年前，R. Watts 和 A. Porter[49] 便提出了以统计关键词变化的方式来探索技术主题的演化，虽然并未考虑词对之间更深层的语义关系，但为后续主题演化分析奠定了基础。针对关键词不能揭示技术主题之间的关联关系这一问题，基于引文的演化分析方法采用测度对象之间的相互引用信息来探测领域的技术主题及演化趋势[50]。但以引文为核心的分析方法无法真正从语义内容的角度深入分析技术的演化和变革，近年来基于科技文本挖掘的主题演化研究受到越来越多研究人员的关注。

1.3　突破性创新主题识别方法

现有突破性创新主题识别主要围绕着文献计量、文本挖掘和网络分析三个视角展开。文献计量视角下，基于引文分析的突破性创新主题识别方法以引文表征技术创新的知识组成，并以引用关系表征文献之间的知识转移。研究人员主要通过专利被引数量[9]、专利科学引文数量[51]、引文新颖性[10]、引文曲线特征[25]、引文关键词或共词网络等[33, 52] 构建相关指标识别突破性创新。此外，该视角下，共词分析相关方法利用词语的频率追踪学科或技术领域的主题变化，也在已有研究中被用来识别突破性创新。例如，J. Kleinberg 等通过不同时间段内词频变化率确定各阶段突发词，进而通过突发词状态变化探寻新的研究热点[32]。近年来，文本挖掘与网络分析相关方法也在突破性创新主题识别领域广受关注，研究人员借助于自然语言处理技术对科技文献和专利文献中的关键词或主题进行挖掘分析，同时依托引文网络发挥复杂网络理论与方法的优势，实现对突破性创新主题的识别与探测[53]。例如，J. Yoon 等基于 SAO 计算专利文本相似度寻找离群专利，用其来表征突破性技术创新[11]。N. Shibata 等把论文的引用网络视为一种复杂网络，对比分析氮化镓（GaN）和复杂网络两个领域，并在引用网络聚类基础上，通过节点的模块内部度和参与系数识别两个领域中的渐进性创新和分岔创新（突破性创新的一种）[12]。

总体来说，虽然文献计量学方法高效、直接，但存在引文时滞性、未深入语义层面、尚停留在对主题的静态文献计量特征进行描述性研究等相关问题。文本挖掘在探究隐含的语义关系方面有一定优势，但不论是以关键词为核心还是基于主题模型的主题抽取方式，在计算主题相似度方面都存在固有问题，很难直接测度技术演化并进行与演化相关的特征分析。最后，复杂网络相关指标的引入为突破性创新识别研究拓展了拓扑性质测度的新视角，不论是引文网络或者语义网络，都可以通过展现知识结构来更好识别突破性创新。基于已有研究，本文综合运用文本挖掘视角与网络分析视角，并将主题演化纳入方法体系，考虑时间因素展示主题的动态变化，在深入挖掘知识结构特征的基础上，综合突破性创新多项特征进行主题识别。

2　研究方法

本文同时运用主题模型、词嵌入、复杂网络分析等

方法，基于动态主题网络上的主题演化及知识结构变动，构建起“新颖性”“突变性”“影响力”和“学科交叉性”的层次指标体系，对突破性创新主题进行识别，整体研究框架如图 1 所示。本文以多个时间窗口下的科研论文数据为数据源，在数据预处理后，首先通过主题模型提取不同时间窗口下的主题集合，并利用 Word2vec 将其映射到统一的向量空间中，生成不同时间窗口下的主题向量矩阵；之后，在网络视角下，本文定义不同时间窗口下主题的演化状态，反映不同主题随时间推移而新生、演化、融合以及消亡的过程；最后，在分析动态主题网络的结构特性变化及知识流动的基础上，面向突破性创新的“新颖性”“突变性”“影响力”及“学科交叉性”构建起层次指标体系，识别突破性创新主题。

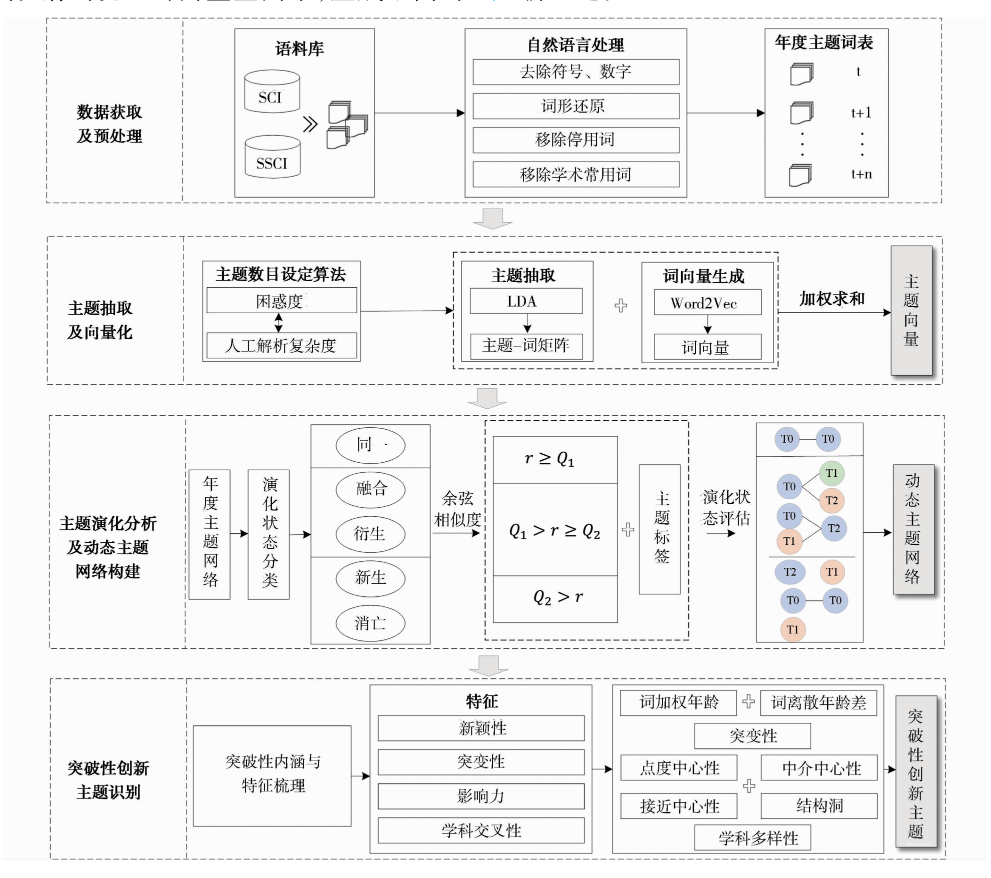


图 1 基于动态主题网络的突破性创新主题识别研究框架

2.1 主题抽取及向量化

2.1.1 基于 LDA 的主题抽取

LDA 是目前应用最为广泛的主题模型之一[47]，其通过随机生成过程来抽取文本在主题空间中的分布，并以词分布的形式表达主题概念。一般来说，可以用随机变量的联合分布来表示 LDA 的生成过程[54]，如公式(1)所示：

$$p(\vec{w}_d,\vec{z}_d,\vec{\theta}_d,\Phi \mid \vec{\alpha},\vec{\beta}) = \prod_{n=1}^{N_d} p(w_{d,n}) \mid \vec{\varphi}_{z_{d,n}})p(z_{d,n} \mid \vec{\theta}_d)p(\vec{\theta}_d \mid \vec{\alpha})\ p(\Phi \mid \vec{\beta}) \quad \text{公式(1)}$$

其中，$\vec{Z}_d$ 是文档 d 的主题分布，$\vec{\theta}_d$ 是对应的主题比例，$Z_{d,n}$代表第 d 个文档中第 n 个单词的主题分布，$\vec{\varphi}_{1:k}$表示主题，每个 $\vec{\varphi}_k$ 是一个词分布，总计 K 个主题，α 和 β 是两个超参数，通常选取默认值[55]。本文将数据集按年份 T 划分，在多个时间窗口上分别训练 LDA 模型，得到 T 个文档 - 主题概率分布矩阵及主题 - 词概率分布矩阵。此外，在主题提取后，采用词分布概率排序及人工校验的方法对所有提取的主题进行命名，生成主题标签，为后续演化状态评估打下基础。

通过 LDA 主题模型进行主题提取需要预先设定主题数目 K，根据已有研究，本文采用综合困惑度[48]及人工解析复杂度[22]双重因素的方法对 K 值进行设定，如公式(2)所示，其中：$Perplexity(D)$ 表示模型的困惑度，计算方法如公式(3)所示；$Complexity$ 表示模型结果的解析复杂度，计算方法如公式(4)所示。困惑度的值越小表明模型与数据的拟合程度越好，而较小的人工解析复杂度则表明解析主题内容的复杂度相对较低，本文综合考虑模型效果与人工解析复杂度确定主题数量。

$$\arg\min_K f(K)=\frac{perplexity(K)-\min perplexity(K)}{\max perplexity(K)-\min perplexity(K)}+\frac{complexity(K)-\min complexity(K)}{\max complexity(K)-\min complexity(K)} \quad 公式(2)$$

$$Perplexity(D)=\exp\left(-\frac{\sum_{d=1}^{M}\log(p(w))}{\sum_{d=1}^{M}N_d}\right) \quad 公式(3)$$

$$Complexity=\exp\left(\frac{K-\min(K)}{\max(K)-\min(K)}\right) \quad 公式(4)$$

在上述困惑度计算公式中，$\sum\log(p(w))$表示给定训练模型在测试语料库上的似然性。

2.1.2　基于 Word2vec 的主题向量化

为了更有效地计算主题相似度并构建动态主题网络，本文运用 Word2vec 算法对主题进行向量化。作为一种高效的词嵌入技术，Word2vec 能够捕捉文本中词语的上下文语境信息，并将词语转化为包含语义关系的低维稠密实数词向量[20]。Word2vec 的具体实现模型包括连续词袋（Continuous Bag-of-word Model）与 skip-gram 模型，根据已有研究，二者在训练效果上不存在显著差异[56]。本文利用 skip-gram 模型将词汇映射为向量，结合主题发现的结果实现文本数据的语义信息提取与向量化，为主题网络及指标识别模型的构建夯实基础。

具体来说，给定文档集合序列 D，其中共包含 N 个单词，N'个非重复的单词。通过 skip-gram 模型对文本集进行训练，生成维度为 γ 的词向量集合 V。由于每个已提取的主题具体表现为 $\vec{\varphi}_k$ 上概率最大的 n 个非重复单词，以每个词对应的概率作为权重，将隶属于该主题非重复单词 γ 维词向量进行加权平均，即可生成统一向量空间中该主题的向量 $v(T_{t,i})$，计算方法如公式(5)所示：

$$v(T_{t,i})=\sum_{j=1}^{s}P(term_{t,i,j})v(term_{t,i,j}) \quad 公式(5)$$

其中，$v(T_{t,i})$为时间窗口 t 下第 i 个主题的向量表示，s 为该主题下概率排名前 s 的词语数，$P(term_{t,i,j})$表示词语对应的概率值，$v(term_{t,i,j})$表示该词语对应的词向量。

2.2　动态主题网络构建及主题演化分析

2.2.1　主题网络构建

以前文生成主题向量为基础，本文利用余弦相似度计算同一时间窗口内主题之间的语义相关程度，并基于主题节点间相似度均值保留连边，构建每个时间窗口内的主题网络，具体计算如公式(6)所示：

$$Similarity_{T_{t,i},T_{t,j}}=\cos(v(T_{t,i}),v(T_{t,j}))=\frac{v(T_{t,i})v(T_{t,j})}{\|v(T_{t,i})\|\cdot\|v(T_{t,j})\|} \quad 公式(6)$$

其中，$T_{t,i}$，$T_{t,j}$表示主题，$v(T_{t,i})$，$v(T_{t,j})$表示主题$T_{t,i}$、$T_{t,j}$的向量形式，$Similarity_{T_{t,i},T_{t,j}}$表示主题 $v(T_{t,i})$、$v(T_{t,j})$之间的相似度，取值介于 0 和 1 之间。该网络的节点是 LDA 抽取的主题，而每一个主题则由带有概率分布的词簇来表示。

2.2.2　主题演化状态界定

整体来看，识别突破性创新主题，需要了解其产生前后的知识状态[13]，而主题演化分析可以揭示科技创新过程中主题的新生、融合、演化、消亡的宏观过程，从而为突破性创新主题识别提供动态视角。在一段时间内发表的某个领域内的科研文献，可以被视作一个随时间延续而发展的动态数据集合。在此数据集上，主题内容演化关系通常表现为某领域内的主题是否出现过，何时出现，与其他哪些主题有关联，关系的发展如何，即是否新出现，或同其他主题合并，亦或是已经消失。以 Y. Zhang 等[57]的研究为基础，本文将主题随时间窗口推移产生的演化状态设为 5 类，即新生、同一、衍生、融合和消亡。各个状态的具体定义如下：

（1）新生主题：新出现的主题，没有任何的承前主题，与先前时间窗口内的主题仅存在较低或者零相关性。

（2）同一主题：现有主题与后续主题关联性极高，两者的相似性达到阈值之上，两者被视为同一个主题。

（3）衍生主题：从现有主题衍化而出的新主题，与当前主题存在较高的相关性但并不十分相似，不属于同一主题，可能存在一对多的关系。

（4）融合主题：融合主题与多个前置主题都有一定的相关性，是多个主题共同融合的结果，但与每个主题都不十分相似，不属于同一主题。

（5）消亡主题：后续时间窗口中所生成的主题与现有主题均不存在相关性，或者相关性极低，则该现有主题可被视为消亡主题。

以上 5 种主题演化状态示意如图 2 所示：

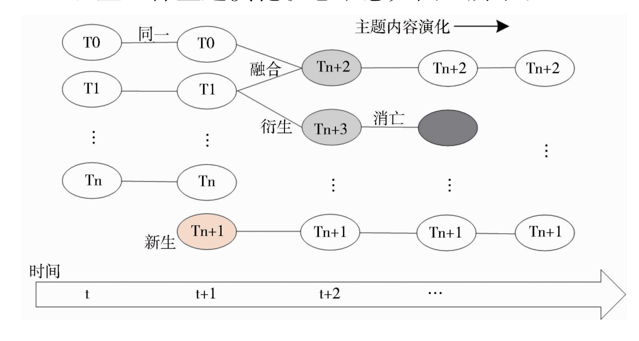


图 2　主题演化状态示意

2.2.3 演化状态测度

合理计算主题相似性是识别主题演化状态变化的基础。为了追踪主题在不同时间窗口上的动态变化与发展，需要捕捉动态数据集每个时间窗口内对应的主题集，并计算相邻窗口间主题集的相似度，从而获悉每个主题的“来源”与“去路”。因此，在识别出每个时间窗口上的主题及其对应的主题向量后，本文采取余弦相似度计算相邻时间窗口主题在语义层面的相关度，如公式(7)所示：

$$Similarity_{T_{t,i},T_{t+1,j}} = cos(v(T_{t,i}), v(T_{t+1,j})) = \frac{v(T_{t,i})v(T_{t+1,j})}{\| v(T_{t,i}) \| \cdot \| v(T_{t+1,j}) \|} \quad 公式(7)$$

其中，$T_{t,i}$ 表示 t 时刻主题，$T_{t+1,j}$ 表示 $t+1$ 时刻主题，$v(T_{t,i})$、$v(T_{t+1,j})$ 分别表示为主题 $T_{t,i}$、$T_{t+1,j}$ 的向量形式，$Similarity_{T_{t,i},T_{t+1,j}}$ 表示主题 $v(T_{t,i})$、$v(T_{t+1,j})$ 之间的相似度，取值介于 0 和 1 之间。

本文结合语义相关度及 2.1.1 小节生成的主题标签，对主题演化状态进行定量测度。如图 3 所示，首先计算每两个相邻时间窗口的主题相似度矩阵（以下简称“相似度矩阵”）的上四分位数（Q_1）与中位数（Q_2）作为主题状态临界点：当主题间相似度达到 Q_1 以上且主题标签相同时，则视两个不同时间窗口下的主题为“同一”主题；若符合 Q_1 阈值条件但未满足主题标签要求，二者之间存在强相关关系，视为衍生或融合状态；若相似度值介于 Q_1 与 Q_2 之间，亦为衍生或融合状态；小于 Q_2 则为新生或者消亡状态。

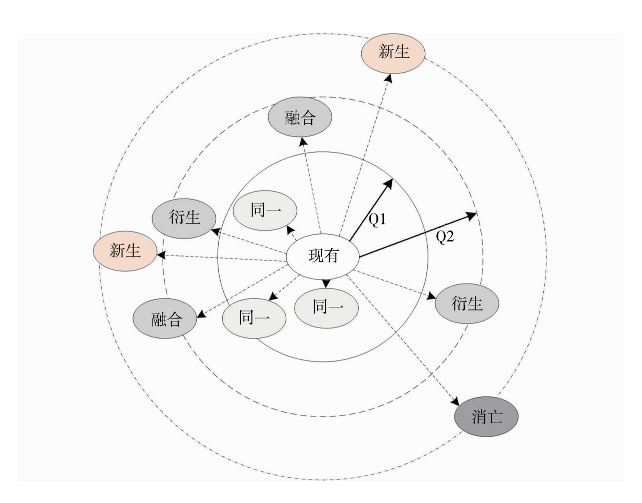


图 3 演化状态界定示意

通过演化状态定量评估，本文依照演化关系将相邻时间窗口上的主题网络进行联通，即以主题演化关系串联起多个时间窗口上的主题网络，形成动态主题网络 G。G 可以表示为：$G=(G_1,G_2,\cdots,G_t,\cdots,G_T)$，其中 G_t 为时间段 t 内的主题网络，T 为划分的时间窗口数量。

2.3 指标构建

本文通过对突破性创新内涵与特征的梳理，构建起测度主题“新颖性”“突变性”“影响力”和“学科交叉性”的层次指标体系，基于已建立的动态主题网络，全面考量主题在时间窗口内的结构特性以及在时间窗口间的演化特征，从而对突破性创新主题进行识别，具体的层次指标体系如表 2 所示。本文在识别突破性创新主题时，不考虑在演化过程中的消亡主题。

表 2 层次指标体系

特征	指标	指标内涵或表征意义
新颖性	主题词语离散年龄差	主题年龄与所有词语平均年龄的离散差
	主题词语加权年龄	主题包含词语的加权平均年龄，即整个数据集中词最早出现年份到当前时刻的年龄差，权重为词分布的概率值
突变性	突变性	相邻时刻主题内容的变化程度，变化程度越大，主题突变程度越高
影响力	点度中心性	与该节点关联的连边数或连边值的总和，表明节点在网络中与其他节点的连接程度
	接近中心性	把一个节点与网络内其它节点连接起来的绝大部分直接路径都是短的（而不是长的）的程度，体现了节点在网络中所占据的重要位置
	中介中心性	网络中经过某个节点的最短路径数目占最短路径总数的比例，反映节点在网络中的影响力
	结构洞	两个节点之间的非冗余的联系，结构洞能够为其占据者获取“信息利益”和“控制利益”提供机会，因而比网络中的其他成员更具优势
学科交叉性	学科多样性	不同学科的交叉往往孕育着新的科学前沿与重大科学突破，主题包含的学科种类越多，跨学科程度越高

2.3.1 基于词语年龄的新颖性测度

在测度技术主题的新颖性时，本文以主题中词语的年龄为核心展开测度，一个词出现的时间越晚，年龄越小，其新颖程度越高。具体地，结合已建立起的动态主题网络，采用词语离散年龄差[58]及词语加权年龄两个指标进行主题新颖度的定量判断，并通过熵权法对其进行拟合，得到综合的新颖性值（Topic Novelty）。词语离散年龄差 Topic Discrete Age Difference（TDA）[58]的具体计算如公式(8)所示：

$$TDA_i = \frac{\sum_{T_i}\left(\frac{1}{n}\sum_{i=n} Y_i - \frac{1}{N}\sum_{j=n} Y_j\right)}{T_i} \quad 公式(8)$$

其中，由于主题 i 在多个时间窗口上动态变化，T_i 表示该主题从新生到最后一个时间窗口所历经的年份总数（时间窗口总数），N 为集合中词的总数，n 表示主题 i 下概率排名靠前的词数量，Y_j 为在整个数据集中，词语出现的最早年份。TDA 有正负之分，值越大，主题中主要词语的出现时间越晚，主题新颖度越高。

词语加权年龄 Topic Weighted Age（TWA）则以测度主题中主要词语的年龄为基础，用词概率加权求和主题中主要词汇从最早出现年份到当前时刻的年龄差，具体计算如公式(9)所示：

$$TWA_i = \frac{\sum_{T_i}\left(\sum_{j=1}^{n} P(w_n)\ Y_j\right)^{-1}}{T_i} \quad 公式(9)$$

其中，T_i 表示主题 i 在演化过程中经历的时间窗口总数，n 为主题 i 下概率排名靠前的词数量，$P(w_n)$ 为对应词的概率值，Y_j 为词 j 的年龄，即整个数据集中该词语最早出现的年份到当前时刻的时间差。TWA 值越大，主题新颖度越高。

2.3.2 基于主题相似性的突变性测度

以前文计算得出的主题向量为基础，本部分基于突变理论设置主题突变度指标（Topic Mutation），主题突变程度越高，表示该主题突变程度越大，反之则越小，具体计算如公式(10)所示：

$$TM_i = \sum_{T_i}\left(1 - \frac{v(T_{t,i}) \cdot v(T_{t+1,i})}{||v(T_{t,i})|| \cdot ||v(T_{t+1,i})||}\right) / T_i \quad 公式(10)$$

其中，T_i 仍然表示主题 i 在演化过程中经历的时间窗口总数，$v(T_{t,i})$、$v(T_{t+1,i})$ 分别为主题 i 在相邻时间窗口上主题向量，该值越大，表示主题的变化程度越高。

2.3.3 基于网络指标的影响力测度

在网络分析中，网络中心性常被用于度量节点在网络中的影响力，相关指标包括接近中心性、介数中心性以及度中心性等[59]，而结构洞常被用于衡量节点的关键位置，本文选取中心性及结构洞来分别测度主题影响力，并通过熵权法进行拟合，得到综合性主题影响力值（Topic Influence）。本小节基于 2.2.1 构建的主题网络进行指标计算，具体如下：

(1)点度中心性可通过主题网络中与主题节点 i 相连的边数与同节点 i 可能相连最大边数之比进行计算，如公式(11)所示[60]：

$$C_D(i) = \frac{\sum_{T_i} k_i/(N-1)}{T_i} \quad 公式(11)$$

其中，T_i 表示主题 i 在演化过程中经历的时间窗口总数，k_i 表示时间窗口 t 下与主题 i 有连边的主题数量，该值越大，表示主题的影响力越高。

(2)接近中心性可通过主题网络中主题节点 i 到其他所有主题节点最短路径的平均长度进行计算，如公式(12)所示[60]：

$$C_C(i) = \frac{\sum_{T_i}(N-1)/\sum_{j\neq i}^{N} d_{ij})}{T_i} \quad 公式(12)$$

其中，d_{ij} 表示主题 i 到主题 j 的最短距离，T_i 表示主题 i 经历的时间窗口总数，该值越大，表明节点位于网络中心位置的程度越大，表示主题的影响力越高。

(3)中介中心性表现为主题网络中经过某个主题节点的最短路径数目占最短路径总数的比例，如公式(13)所示[60]：

$$C_B(i) = \frac{\sum_{T_i}\sum_{j\neq i\neq k\in V, j<k}\sigma_{jk}(i)/\sigma_{jk}}{T_i} \quad 公式(13)$$

其中，T_i 代表主题 i 在演化过程中经历的时间窗口总数，$\sigma_{jk}(i)$ 为时间窗口 t 下，节点 j 与 k 之间最短路径通过节点 i 的数目，σ_{jk} 为节点 j 与 k 之间所有最短路径的总数，该指标反映节点在网络中的影响力，值越大，表示主题的影响力越高。

(4)结构洞相关研究中，通常以网络约束系数来计算各节点所占有的位置优势，以其描绘某节点与其他节点直接或间接联系的紧密程度，该值越小，结构洞越多，位置越重要，该节点越具有获取多样化知识的能力，是潜在的创新节点[61]。本文利用 UCINET 软件计算主题 i 在演化过程中，每个时间窗口上，主题网络中结构洞约束系数，并以 i 在演化过程中经历的时间窗口总数进行加和平均，最终得到多个窗口下主题 i 的结构洞约束系数。

2.3.4 基于学科分类的学科交叉性测度

不同学科的交叉点往往是新科学的生长点和新的科学前沿，也最有可能产生重大科学突破[62]。根据已有研究，本部分基于 Web of Science 的学科分类（Web of Science Category）来计算动态网络上主题的学科交叉程度。由于每篇论文隶属于一个或多个学科，同时涵盖若干个主题，即每个主题下不同学科的贡献程度有所差异，本文提出学科多样性（Topic Subject Diversity）指标，用以表征主题的学科交叉性。该值越大，表示主题包含的学科种类越多，跨学科程度越高，计算方法如公式(14)所示：

$$TST_i = \sum_{T_i}\frac{\sum_{m=1}^{h} P(d_{t,i,m})\ S(d_{t,i,m})}{S\sum_{m=1}^{h} P(d_{t,i,m}))} / T_i \quad 公式(14)$$

其中，T_i 表示主题 i 在演化过程中经历的时间窗

口数量,h 表示时间窗口 t 下文档数量,$P(d_{t,i,m})$ 表示时间窗口 t 中第 i 个主题下第 m 个文档属于该主题的概率,$S(d_{t,i,m})$ 表示时间窗口 t 下第 m 个文档包含的学科数量,S 表示 WoS 中的学科分类总数。

3 实证分析

区块链作为当前信息技术的前沿领域之一,以其基础性、引领性和创新性等特征,不断激发、赋能和提速数字经济发展,对当前的信息技术形成了全方位、战略性影响。下文以区块链领域为例,对其相关科研论文数据进行较为全面的采集,开展基于动态主题网络的突破性创新主题识别,验证本文方法及相关研究工作的可行性及有效性。

3.1 数据获取

区块链作为高速发展的前沿领域,有关其检索策略尚未达成共识,现有研究大多以"区块链"或"blockchain"或"bitcoin"为关键词进行中英文文献检索。本文通过梳理区块链领域相关文献,在商琦和陈洪梅[63]构建的检索策略基础上进行改进,得到检索策略如下:TS = (" chain of block * " OR " blockchain * " OR " block chain * " OR " genesis block * " OR " Bitcoin" OR " Ethereum" OR " Consensus mechanism * " OR " proof of work * " OR " proof-of-work * " OR " proof of stake * " OR " proof-of-stake * " OR " Byzantine Fault Toleran * " OR " Proof of Authority" OR " Proof-of-Authority" OR " Distributed ledger * " OR " smart contract * " OR " asymmetric encryption * ")。应用以上检索式,本文在 Web of Science 的 SCI 以及 SSCI 数据库中,检索 2011 – 2020 年的英文期刊与会议文献,总计获得 10 817 条数据。为进一步提高数据准确性,对下载文献通过人工干预移除少量化学、材料学、免疫学、细胞学和药学以及其他与区块链核心内容相关性较弱的数据条目[63],筛查之后保留区块链相关数据 9 805 条,形成该领域突破性创新主题识别的初始语料库。图 4 展示了区块链领域论文的年度发表数量。可以看出,前期相关文献数量较少,且增速相对平缓,自 2016 年文献数量开始大幅增加。

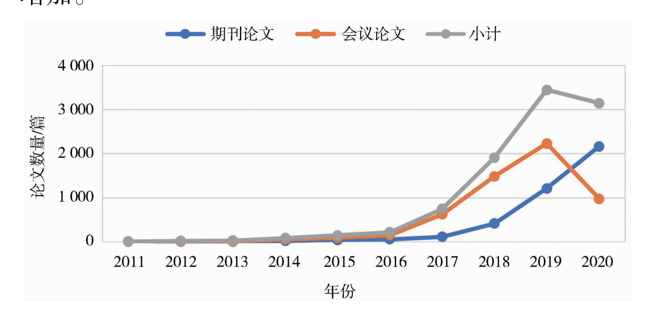


图 4 年度文献数量

3.2 动态主题网络构建

为了仅保留和区块链技术最为直接相关的文本内容,本文通过自然语言处理对初始语料库中的标题和摘要字段进行清洗,移除停词(Stopwords)及常用语等内容。而后,以年为单位,建立 2011 年至 2020 年 10 个时间窗口,并按照时间窗口划分文本集。对于每个窗口下的文本集,本文平衡困惑度及人工解析复杂度,将 10 个阶段的主题数量 K 设置为:7、10、12、14、15、25、25、25、30 及 25。因篇幅原因,本文仅展示 2020 年时间窗口下主题总数参数的确定过程,其余时间窗口下计算逻辑相同,如图 5 所示:

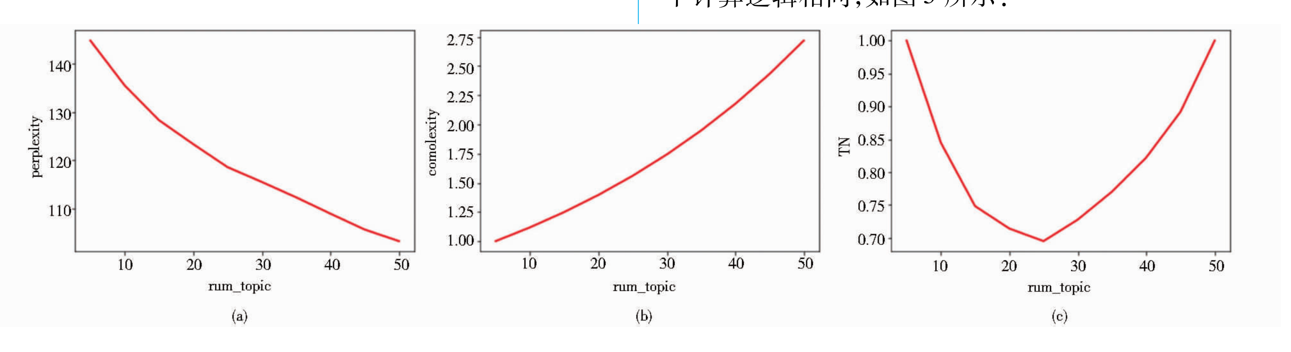


图 5 主题数量确定流程(以 2020 年文本集为例)

而后,本文采用 2 000 次 Gibbs 采样迭代来推断潜在变量和分布,以提取每个时期的主题,并采用词分布概率排序及人工校验的方法对各个网络中的主题进行命名,生成主题标签;同时,使用 Python Genism 工具包在整体语料库上训练词向量,维度参数 γ 设置为 150,窗口大小设置为 5。根据 2.1.2 给出的方法,每个时间窗口下的主题都被转化成了统一向量空间内维度为 150 维的向量。本文通过公式(6)计算主题相关矩阵,构建各时期的主题网络,如图 6 所示(由于篇幅限制,这里仅展示时间窗口 1 与 10 对应的主题网络)。

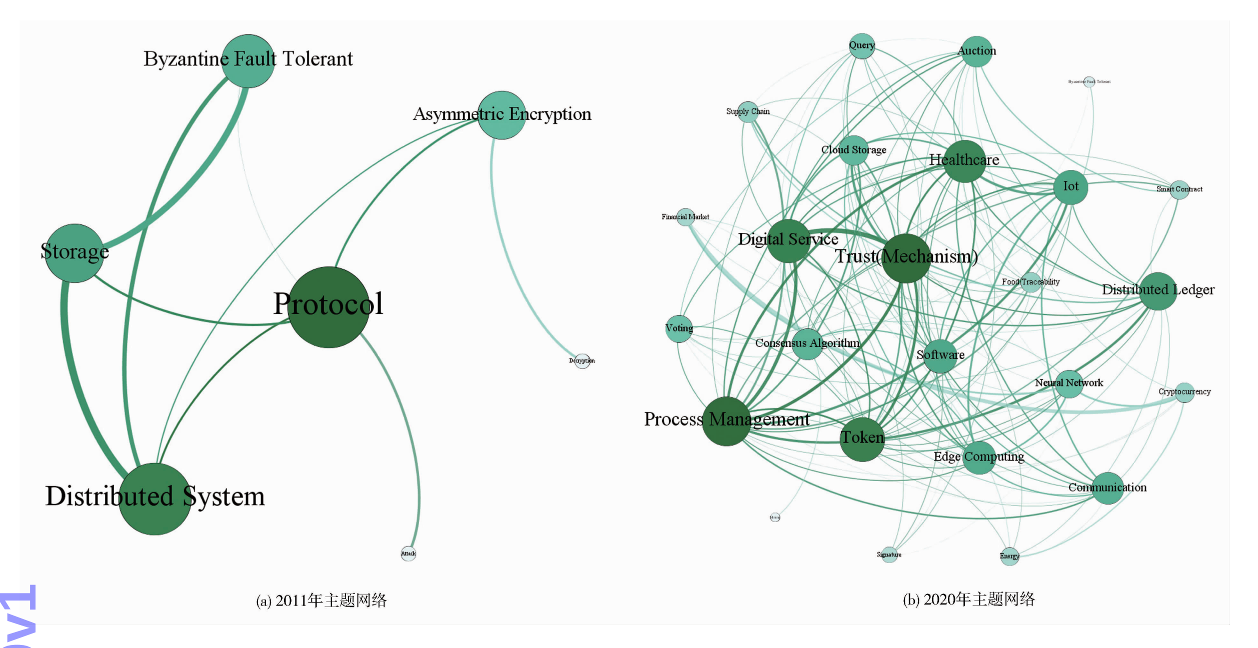


图 6　主题网络

为了追踪主题在不同时间窗口下的动态变化与发展，本文随后计算相邻时间窗口间主题集的相似度，识别区块链领域内每个主题的“来源”与“去路”。基于已经计算得出的主题向量，本部分采用余弦相似度共得出 9 个表示相邻时间窗口主题集合变化的相关性矩阵，并通过计算上四分位数(Q_1)与中位数(Q_2)对主题在窗口间的演化状态进行定量评估。基于演化状态测度结果，在过去的 10 年中，区块链领域共出现了 87 个不重复的主题，涵盖所有的新生、同一、衍生、融合和消亡状态，它们的演化过程如图 7 所示。其中，实线表示同一主题，虚线表示衍生或者融合状态，深灰色节点表示消亡主题。从图 7 中可以看出，从 2016 年开始主题数量增多，2019 年演化出更多研究主题，大多主题都出现了衍生或者融合状态，各主题处于动态变化之中，衍生或者融合演化是该领域内知识流动的常态。

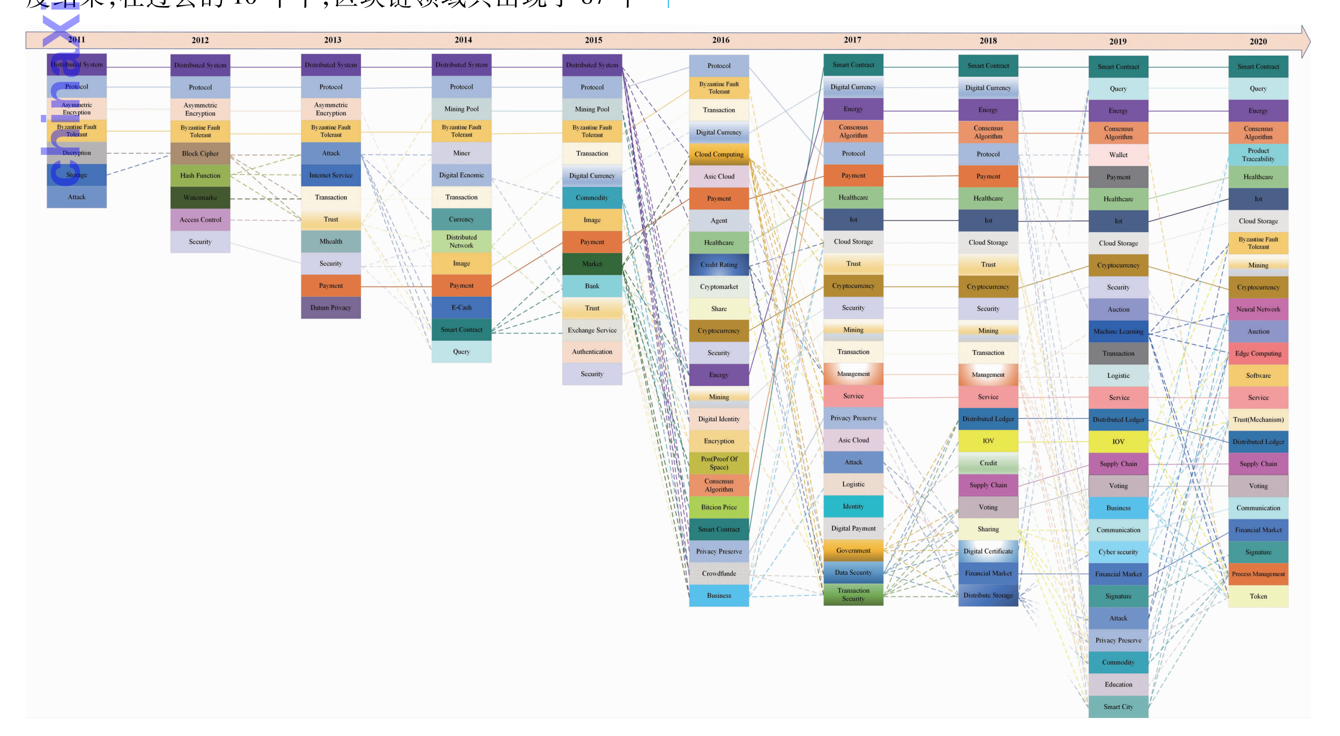


图 7　2011－2020 年区块链领域主题演化情况

3.3　突破性创新主题识别

基于动态主题网络,运用 2.3 小节中的指标计算方法,可得到每个主题的“新颖性”“突变性”“影响力”和“学科交叉性”指标值,并进行标准化处理,将其统一映射到[0,1]区间内,如表 3 所示。根据各指标均值,本文将 2020 年现存的 25 个主题(见表 4)划分为 16 个类别,分别表征各主题在 4 个维度指标上的整体特性。其中有 2 个主题属于新颖性高、突变性高、影响力大、学科交叉性强的类别,被识别为突破性创新主题,分别为神经网络(2020 – T6 – Neural Network)和边缘计算(2020 – T8 – Edge Computing)。

表 3　2020 年 25 个主题“新颖性”“突变性”“影响力”和“学科交叉性”指标值

主题	新颖性	突变性	影响力	学科交叉性	主题	新颖性	突变性	影响力	学科交叉性
Distributed Ledger	0.151 8	0.959 3	0.701 2	0.798 3	Cloud Storage	0.044 2	0.578 8	0.716 4	0.814 5
Product Traceability	0.148 7	0.936 5	0.484 2	0.788 3	Mining	0.100 8	0.880 5	0.404 0	0.805 2
Byzantine Fault Tolerant	0.068 7	0.900 1	0.507 1	0.807 7	IoT	0.098 6	0.071 4	0.668 1	0.867 8
Communication	0.059 0	0.752 5	0.783 0	0.814 5	Process Management	0.087 7	0.851 6	0.727 9	0.831 7
Cryptocurrency	0.313 6	0.773 9	0.108 4	0.606 4	Voting	0.229 7	0.578 5	0.590 5	0.825 8
Consensus Algorithm	0.055 1	0.791 2	0.522 1	0.843 1	Query	0.042 9	0.710 2	0.533 6	0.795 5
Neural Network	0.511 5	1.000 0	0.639 8	0.799 0	Smart Contract	0.039 9	0.055 7	0.292 8	0.806 4
Financial Market	0.273 4	0.752 5	0.606 6	0.664 3	Energy	0.104 2	0.123 1	0.305 6	0.000 0
Edge Computing	0.269 1	0.841 3	0.678 9	0.844 5	Healthcare	0.197 4	0.678 2	0.351 5	0.820 1
Digital Service	0.047 5	0.674 2	0.950 1	0.864 5	Token	0.084 1	0.901 8	0.735 4	0.818 3
Supply Chain	0.952 5	0.000 0	0.025 7	1.000 0	Auction	0.068 5	0.207 7	0.850 3	0.871 8
Software	0.091 6	0.719 6	0.708 1	0.816 5	Signature	0.046 4	0.696 1	0.533 4	0.797 8
Trust (Mechanism)	0.052 5	0.797 3	0.760 9	0.863 8	均值	0.165 6	0.649 3	0.567 4	0.782 6

表 4　2020 年主题的标签、主要内容与突破性创新特征测度

全局标号	局部标号	主题标签	主要内容	新颖性	突变性	影响力	学科交叉性
T67 – 2018	2020 – T0	Distributed Ledger	distribute_ledger, payment, scalability, DLT, application	低	高	高	高
T90 – 2020	2020 – T1	Product Traceability	product, industry, supply_chain, food, traceability	低	高	低	高
T91 – 2020	2020 – T2	Byzantine Fault Tolerant	event, byzantine fault tolerant (BFT), attack, consensus, byzantine	低	高	低	高
T80 – 2019	2020 – T3	Communication	communication, terminal, module, blockchain-base, efficiency	低	高	高	高
T50 – 2016	2020 – T4	Cryptocurrency	market, cryptocurrencie, price, return, cryptocurrency	高	高	低	低
T46 – 2016	2020 – T5	Consensus Algorithm	node, block, consensus, protocol, consensus_algorithm	低	高	低	高
T92 – 2020	2020 – T6	Neural Network	ICO, neural_network, algorithm, neuron, computing_power	高	高	高	高
T74 – 2018	2020 – T7	Financial Market	money, financial_market, bitcoin, USD, price_movement	高	高	高	低
T93 – 2020	2020 – T8	Edge Computing	edge, algorithm, edge_computing, mobile, AI	高	高	高	高
T59 – 2017	2020 – T9	Digital Service	digital, trust, service, agent, blockchain-base	低	高	高	高
T70 – 2018	2020 – T10	Supply Chain	adoption, supply_chain, SC (supply chain), management, tourism	高	低	低	高
T94 – 2020	2020 – T11	Software	attack, SDN (Software defined network), distribute, software, transmission	低	高	高	高
T95 – 2020	2020 – T12	Trust(Mechanism)	trust, service, content, user, mechanism	低	高	高	高
T60 – 2017	2020 – T13	Cloud Storage	cloud, service, user, image, cloud_storage	低	低	高	高
T94 – 2020	2020 – T14	Mining	miner, mining, game, mining_pool, revenue	低	高	低	高
T66 – 2017	2020 – T15	IoT	IoT, device, security, smart, IoT_device	低	低	高	高
T99 – 2020	2020 – T16	Process Management	application, process, management, service, industry	低	高	高	高
T71 – 2018	2020 – T17	Voting	block_chain, voting, credit, process, vote	高	低	高	高

（续表4）

全局标号	局部标号	主题标签	主要内容	新颖性	突变性	影响力	学科交叉性
T76 - 2019	2020 - T18	Query	query, security, protocol, set, message	低	高	低	高
T49 - 2016	2020 - T19	Smart Contract	smart_contract, contract, ethereum, execution, cost	低	低	低	高
T41 - 2016	2020 - T20	Energy	energy, trading, power, market, energy_trading	低	低	低	低
T36 - 2016	2020 - T21	Healthcare	security, healthcare, user, medical, storage	高	高	低	高
T98 - 2020	2020 - T22	Token	user, token, social, ethereum, application	低	高	高	高
T86 - 2020	2020 - T23	Auction	insurance, auction, financial, fraud, framework	低	低	高	高
T77 - 2019	2020 - T24	Signature	signature, content, algorithm, digital, key	低	高	低	高

主题6神经网络(2020 - T6 - Neural Network)主要涵盖首次代币发行(ICO)、神经网络、算法、算力等内容,涉及金融领域中神经网络算法与区块链技术的融合应用。从已有区块链相关研究可以验证该主题的突破性创新属性:区块链+神经网络可以有效提高交易的安全性、身份认证的可靠性以及解决信息不公开等问题,能够为实体经济发展和实现数字经济生态提供技术保障和强大驱动力,其取得的突破性进展在近年来得到广泛关注[64-66]。美国国家科学技术委员会(National Science and Technology Council,NSTC)于2022年2月发布了新一版关键和新兴技术(Critical and Emerging Technologies, CETs)清单[67],将分布式记账技术(区块链技术)纳入金融技术类别,充分体现了其重要程度和两个领域的快速融合。

主题8边缘计算(2020 - T8 - Edge Computing)主要涉及边缘计算、移动边缘计算、边缘人工智能计算等内容。已有相关研究显示,边缘计算能够为区块链服务提供资源,主要包括通信资源和计算资源[68],区块链技术负责保障安全,边缘计算负责提高通信效率。2020年,中国移动发布《区块链+边缘计算技术白皮书》,指出"区块链+边缘计算"的融合应用作为通信和信息技术融合发展的新领域,能够促进资源共享、最优化配置以及跨界协同和创新,加快社会信息化转型,研究前景广阔[69]。此外,美国国家科学技术委员会发布的CETs清单将边缘计算列为先进计算的代表,这也在一定程度上验证了本文识别结果的有效性。总体看来,"区块链+"产业融合模式迅速发展,在为各领域带来深刻变革的同时,也标志着区块链发展进入3.0时代[70],开启了全新的发展阶段。

4 总结与展望

突破性创新是国家在产业革命浪潮中把握制胜先机、企业提高竞争力的关键要素。准确地识别突破性创新主题能够为国家政策制定及企业战略布局提供决策支持,为学界聚焦研究重点指明方向。归纳总结已有研究,突破性创新主题需要通过动态视角进行分析,且揭示科研主题的动态演化过程、规律和态势对于突破性创新主题的探测具有至关重要的意义。本文以多个时间窗口下的科研论文数据为数据源,综合运用概率主题模型与词嵌入的方法进行主题的抽取与向量化,首先克服了以关键词为核心的主题识别方式在语义表达上存在盲点和筛选及降维困难等问题,完成了科技文本到数学向量的映射。随后,本文在连续时间窗口下构建起动态主题网络,全面考量主题在时间窗口内的结构特性以及在时间窗口间的演化特征,并构建起测度主题"新颖性""突变性""影响力"和"学科交叉性"的层次指标体系,对突破性创新主题进行识别。从方法上看,动态主题网络下的突破性创新主题识别研究是对现有基于文本挖掘和网络分析视角方法的重要补充。

从结果上看,本文使用2011-2020年区块链领域的科研文献数据,识别出2个突破性创新性质最为显著的主题,即神经网络和边缘计算,并结合该领域已有研究及技术清单验证了方法的有效性。但是,本文存在一定的不足及继续研究的空间。首先,在动态主题网络的视角下,没有构建起定量的结果验证方法;其次,本文目前仅考虑了科研文献数据,数据来源单一,需要在研究中进一步拓展数据维度,并将突破性创新的更多特性映射到多源、异质的动态主题网络之上;最后,本文的方法仅在区块链领域进行了实证分析,未来需要在其他技术领域展开分析,进一步验证方法的系统性和可靠性。

参考文献:

[1] 郭小超，雷婧，冯银虎，等. 基于知识图谱的国际突破性创新理论研究综述[J]. 科学管理研究，2020，38(1)：20-26.
[2] 邵云飞，詹坤，吴言波. 突破性技术创新:理论综述与研究展

望[J]. 技术经济, 2017, 36(4): 30 –37.
[3] HAIN D S, CHRISTENSEN J L. Capital market penalties to radical and incremental innovation[J]. European journal of innovation management, 2020, 23(2): 291 –313.
[4] 万宁. 浅析颠覆性创新、破坏性创新和突破性创新三者关系[J]. 商, 2015(30): 122 –123.
[5] ABERNATHY W J, UTTERBACK J M J T R. Patterns of innovation in technology[J]. Technology review, 1978, 80(7): 41 –47.
[6] MCDERMOTT C M, O'CONNOR G C. Managing radical innovation: an overview of emergent strategy issues[J]. Journal of product innovation management, 2002, 19(6): 424 –438.
[7] LEIFER R. Radical innovation: how mature companies can outsmarts upstatrs[M]. Brighton: Harvard Business Press, 2000.
[8] 张金柱, 张晓林. 基于专利科学引文的突破性创新识别研究述评[J]. 情报学报, 2016, 35(9): 955 –962.
[9] SCHOENMAKERS W, DUYSTERS G. The technological origins of radical inventions[J]. Research policy, 2010, 39(8): 1051 –1059.
[10] DAHLIN K B, BEHRENS D M. When is an invention really radical? defining and measuring technological radicalness[J]. Research policy, 2005, 34(5): 717 –737.
[11] YOON J, KIM K. Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks[J]. Scientometrics, 2011, 88(1): 213 –228.
[12] SHIBATA N, KAJIKAWA Y, TAKEDA Y, et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications[J]. Technovation, 2008, 28(11): 758 –775.
[13] 刘亚辉, 许海云. 突破性创新早期识别与弱信号分析综述[J]. 图书情报工作, 2021, 65(4): 89 –101.
[14] ZHANG Y, ZHANG G, CHEN H, et al. Topic analysis and forecasting for science, technology and innovation: methodology with a case study focusing on big data research[J]. Technological forecasting and social change, 2016, 105: 179 –191.
[15] 李慧, 玄洪升. 专利视角下融合多属性的技术创新主题挖掘方法——以芯片领域专利为例[J]. 图书情报工作, 2020, 64(11): 96 –107.
[16] CHEN H, WANG X, PAN S, et al. Identify topic relations in scientific literature using topic modeling[J]. IEEE transactions on engineering management, 2021, 68(5): 1232 –1244.
[17] CHEN H, ZHANG G, ZHU D, et al. Topic-based technological forecasting based on patent data: a case study of Australian patents from 2000 to 2014[J]. Technological forecasting and social change, 2017, 119: 39 –52.
[18] SUOMINEN A, TOIVANEN H. Map of science with topic modeling: comparison of unsupervised learning and human-assigned subject classification[J]. Journal of the Association for Information Science and Technology, 2016, 67(10): 2464 –2476.
[19] JUNG S, YOON W C. An alternative topic model based on common interest authors for topic evolution analysis[J]. Journal of informetrics, 2020, 14(3): 101040.
[20] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th international conference on neural information processing systems. New York: ACM, 2013: 3111 –3119.
[21] ZHANG Y, LU J, LIU F, et al. Does deep learning help topic extraction? a kernel k-means clustering method with word embedding[J]. Journal of informetrics, 2018, 12(4): 1099 –1117.
[22] CHEN H, JIN Q, WANG X, et al. Profiling academic-industrial collaborations in bibliometric-enhanced topic networks: a case study for digitalization research[J]. Technological forecasting and social change, 2022, 175: 121402.
[23] DEWAR R D, DUTTON J E. The adoption of radical and incremental innovations: an empirical analysis[J]. Management science, 1986, 32(11): 1422 –1433.
[24] LI H, ZHANG Q, ZHENG Z. Research on enterprise radical innovation based on machine learning in big data background[J]. Journal of supercomputing, 2020, 76(5): 3283 –3297.
[25] 曹艺文, 许海云, 武华维, 等. 基于引文曲线拟合的新兴技术主题的突破性预测——以干细胞领域为例[J]. 图书情报工作, 2020, 64(5): 100 –113.
[26] ZHOU K Z, YIM C K, TSE D K. The effects of strategic orientations on technology- and market-based breakthrough innovations[J]. Journal of marketing, 2005, 69(2): 42 –60.
[27] 许海云, 刘亚辉, 罗瑞. 突破性科学创新早期识别研究综述[J]. 情报理论与实践, 2021, 44(4): 198 –205.
[28] ARTS S V R. The technological origins and novelty of breakthrough inventions[C]//35th DRUID celebration conference. Spain: Barcelona, 2013: 1 –30.
[29] 周磊, 杨威, 张玉峰. 基于专利挖掘的突破性创新识别框架研究[J]. 情报理论与实践, 2016, 39(9): 73 –76,46.
[30] 黄鲁成, 蒋林杉, 吴菲菲. 萌芽期颠覆性技术识别研究[J]. 科技进步与对策, 2019, 36(1): 10 –17.
[31] PONOMAREV I, LAWTON B K, WILLIAMS D E, et al. Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? [J]. Scientometrics, 2014, 100(3): 755 –765.
[32] KLEINBERG J. Bursty and hierarchical structure in streams[J]. Data mining and knowledge discovery, 2003, 7(4): 373 –397.
[33] 张金柱, 张晓林. 基于被引科学知识主题突变的突破性创新识别[J]. 现代图书情报技术, 2016(Z1): 42 –50.
[34] CHEN C, CHEN Y, HOROWITZ M, et al. Towards an explanatory and computational theory of scientific discovery[J]. Journal of informetrics, 2009, 3(3): 191 –209.
[35] AHUJA G, LAMPERT C M. Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions[J]. Strategic management journal, 2001, 22

(6/7): 521－543.
[36] 张军. 破坏性创新的特征分析[J]. 商场现代化, 2007(27): 76.
[37] 张栋. 面向2035年的突破性创新测度、识别与预测[J]. 中国科技论坛, 2020(8): 11－14.
[38] 庄子银, 贾红静, 肖春唤. 突破性创新研究进展[J]. 经济学动态, 2020(9): 145－160.
[39] DELLA MALVA A, KELCHTERMANS S, LETEN B, et al. Basic science as a prescription for breakthrough inventions in the pharmaceutical industry[J]. Journal of technology transfer, 2015, 40(4): 670－695.
[40] DESS G G P S D. Porter's generic strategies as determinants of strategic group membership and organizational performance[J]. Academy of management journal, 1984, 27(3): 467－488.
[41] DOSI G. Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change[J]. Research policy, 1982, 11(3): 147－162.
[42] 付玉秀, 张洪石. 突破性创新:概念界定与比较[J]. 数量经济技术经济研究, 2004(3): 73－83.
[43] ANDERSON P, TUSHMAN M L. Technological discontinuities and dominant designs: a cyclical model of technological change[J]. Administrative science quarterly, 1990, 35(4): 604－633.
[44] 李良德, 陈劲, 莫昕玮. 突破性创新管理模式研究[J]. 中外科技信息, 2001(11): 38－41.
[45] WANG X, WANG Z, HUANG Y, et al. Identifying R&D partners through subject-action-object semantic analysis in a problem & solution pattern[J]. Technology analysis & strategic management, 2017, 29(10): 1167－1180.
[46] 胡正银, 方曙. 专利文本技术挖掘研究进展综述[J]. 现代图书情报技术, 2014(6): 62－70.
[47] BLEI D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77－84.
[48] DE BATTISTI F, FERRARA A, SALINI S. A decade of research in statistics: a topic model approach[J]. Scientometrics, 2015, 103(2): 413－433.
[49] WATTS R J, PORTER A L. Innovation forecasting[J]. Technological forecasting and social change, 1997, 56(1): 25－47.
[50] 赵蓉英, 郭凤娇, 赵月华. 科学计量学主流研究领域与热点前沿研究[J]. 图书情报工作, 2015, 59(2): 66－74.
[51] TORTORIELLO M, MCEVILY B, KRACKHARDT D. Being a catalyst of innovation: the role of knowledge diversity and network closure[J]. Organization science, 2015, 26(2): 423－438.
[52] 张金柱, 张晓林. 利用引用科学知识突变识别突破性创新[J]. 情报学报, 2014, 33(3): 259－266.
[53] 卢超, 侯海燕, DING Y, 等. 国外新兴研究话题发现研究综述[J]. 情报学报, 2019, 38(1): 97－110.
[54] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101: 5228－5235.
[55] WEI X, CROFT W B. LDA-based document models for ad hoc retrieval[C]//The 29th annual international ACM SIGIR conference on researchand development in information retrieval. New York: Association for Computing Machinery, 2006: 178－185.
[56] LEVY O, GOLDBERG Y, DAGAN I. Improving distributional similarity with lessons learned from word embeddings[J]. Transactions of the Association for Computational Linguistics, 2015, 3: 211－225.
[57] ZHANG Y, ZHANG G Q, ZHU D H, et al. Scientific evolutionary pathways: identifying and visualizing relationships for scientific topics[J]. Journal of the Association for Information Science and Technology, 2017, 68(8): 1925－1939.
[58] 沈君, 王续琨, 高继平, 等. 技术坐标视角下的主题分析——以第三代移动通信技术为例[J]. 情报学报, 2012, 31(6): 603－611.
[59] 姜鑫, 王德庄, 马海群. 社会网络分析方法在图书情报学科的应用研究[M]. 北京: 知识产权出版社, 2019.
[60] KONG X J, SHI Y J, YU S, et al. Academic social networks: modeling, analysis, mining and applications[J]. Journal of network and computer applications, 2019, 132: 86－103.
[61] 宋歌. 网络结构视域下的创新潜力指标研究[J]. 图书情报工作, 2014, 58(3): 64－71.
[62] 路甬祥. 学科交叉与交叉科学的意义[J]. 中国科学院院刊, 2005(1): 58－60.
[63] 商琦, 陈洪梅. 区块链技术创新态势专利情报实证[J]. 情报杂志, 2019, 38(4): 23－28,59.
[64] 闫凯伦. 面向区块链的交易传播算法和去中心化机器学习框架研究[D]. 桂林:广西师范大学, 2021.
[65] 何帅, 黄襄念, 刘谦博, 等. DPoS区块链共识机制的改进研究[J]. 计算机应用研究, 2021, 38(12): 3551－3557.
[66] 朱书坤. 基于区块链和卷积神经网络的电动汽车能源交易方案设计与实现[D]. 武汉:华中师范大学, 2020.
[67] NSTC. Critical and emerging technologies, CETs[R]. Washington, DC: National Science and Technology Council, 2022.
[68] 武继刚, 刘同来, 李境一, 等. 移动边缘计算中的区块链技术研究进展[J]. 计算机工程, 2020, 46(8): 1－13.
[69] 何申, 陆璐, 李征, 等. 区块链＋边缘计算技术白皮书[R]. 杭州: 中国移动5G联合创新中心, 2020.
[70] 王晨旭, 程加成, 桑新欣, 等. 区块链数据隐私保护:研究现状与展望[J]. 计算机研究与发展, 2021, 58(10): 2099－2119.

作者贡献说明：

陈虹枢：方法构思、论文撰写；
宋亚慧：数据获取、论文实验、论文撰写；
金茜茜：方法检验、论文撰写；
汪雪锋：方法检验、论文撰写。

Radical Innovative Topic Identification from a Perspective of Dynamic Topic Network: Taking the Field of Blockchain as an Example

Chen Hongshu Song Yahui Jin Qianqian Wang Xuefeng
School of Management and Economics, Beijing Institute of Technology, Beijing 100081

Abstract: [Purpose/Significance] Radical innovation plays a key role in the development of science and technology. In the big data environment, the complex, multidimensional, and continuous evolutionary characteristics of science and technology development itself is becoming more observable than ever before. It is important to identify these topics from a dynamic perspective to provide solutions for countries, enterprises and universities to analyze radical innovation areas, allocate innovation resources rationally and seek innovation upgrades. [Method/Process] This paper integrated methods of topic modeling, word embedding algorithm, and complex network analysis to construct dynamic topic networks, and evaluate the structural characteristics of the topics within different time windows and the topic evolution states between these time windows. Based on dynamic topic networks, this paper then combined the novelty, mutation, impact and interdisciplinary characteristics of radical innovation to identify topics of radical innovation. [Result/Conclusion] Through the empirical study on blockchain, this paper recognizes that two topics with the most significant radical innovative characteristics are Neural Network and Edge Computing. With existing research of blockchain and the list of critical and emerging technologies issued by the National Science and Technology Council (NSTC) of the United States, this paper finally verifies the feasibility and effectiveness of the proposed method. However, further quantitative verification of the result of this paper, and identification of radical innovative topics by fusing multi-source data, require further research in the future.

Keywords: radical innovation topic network topic identification LDA Word2vec blockchain